# AN IMPROVED ALGORITHM FOR RESOURCE ALLOCATION AND SCHEDULING IN CLOUD COMPUTING

**Synopsis Submitted to the IIMT University in Partial Fulfillment of the Requirements for the Award of the Degree of**

DOCTOR OF PHILOSOPHY

In

Computer Science

By

*(Neeraj Kumar Gupta)*

Full time

Under the guidance of

Dr. Rattan Pal Singh Rana

School of Computer Science and Applications IIMT University, Meerut

IIMT University, Meerut

(School of Computer Science and Applications IIMT University, Meerut)
October-2021

# TABLE OF CONTENTS

1. INTRODUCTION ..... 3
1.1 Cloud Components ..... 5
2. LITERATURE REVIEW ..... 7
2.1 Resource Allocation in Cloud Computing ..... 7
2.2 Significance of Allocating Resources ..... 8
2.3 Approaches to Allocating Resources ..... 8
2.4 Noteworthy Contribution ..... 9
3. PROBLEM STATEMENT ..... 12
4. RESEARCH OBJECTIVE ..... 12
5. RESEARCH METHODOLOGY ..... 12
6. SCOPE OF RESEARCH ..... 13
7. RESEARCH SCHEDULE ..... 15
CONCLUSION ..... 15
REFERENCES ..... 16

**SYNOPSIS**

**Topic:** An improved algorithm for resource allocation and scheduling in cloud computing.

## 1. INTRODUCTION

In accordance with NIST's SP 800-145 [1], cloud computing serves as a framework that enables widespread and convenient network access to miscellaneous configurable computing resources for sharing as applications, servers, networks, storage, , and services that can be quickly allocated and commuted with least service provider interaction or management effort.

The three primary highlights or key aspects of cloud computing according to NIST can be outline in the following manner [2]:

- Software as-a-Service (SaaS)
- Platform-as-a-Service (PaaS)
- Infrastructure-as-a-Service (IaaS)

Cloud service providers (CSPs) utilize these services to serve their customers using four models as per NIST[3]:

- Private Cloud
- Hybrid Cloud
- Community Cloud
- Public Cloud

Apart from these services and models an integrated view of following five essential and unique characteristics that can be found in every cloud service:

- On-demand Self Service
- Resource Pooling
- Measured Services

- Rapid Elasticity
- Broad Network Access

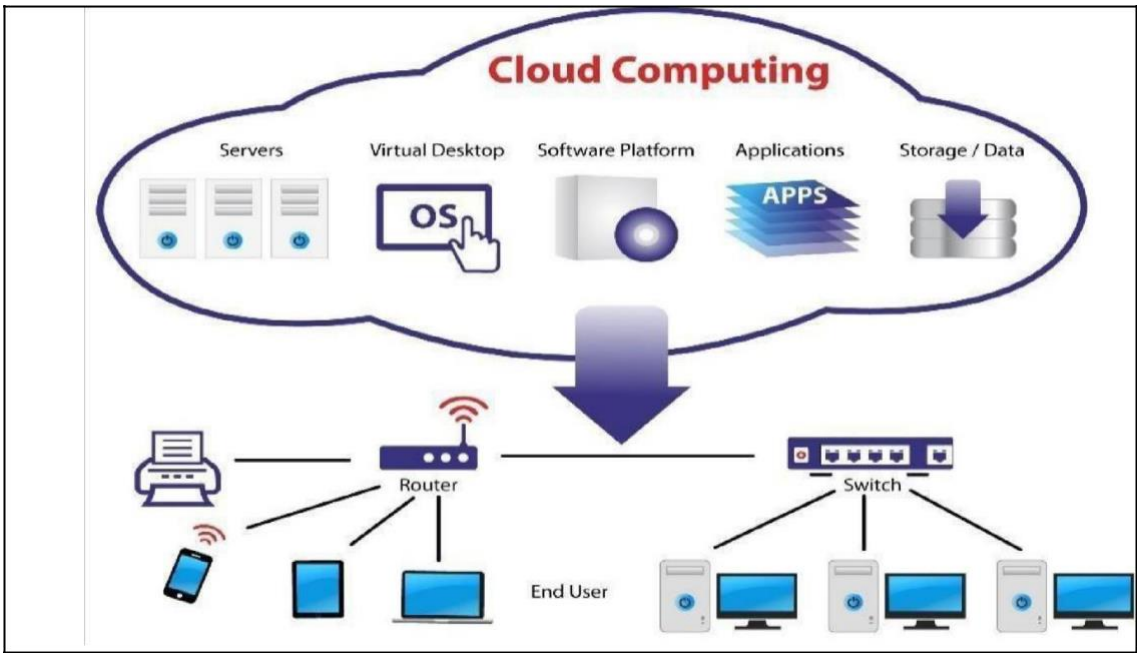


Fig. 1.1: Typical Cloud Computing Environment

Four types of cloud computing deployment models are:

- ***Private Cloud:***A cloud-based processing environment and application can provide clients with precisely the cloud computing model they require for a private cloud setup. [4]
- ***Public Cloud:***As mentioned in [4], the user's computing infrastructure can utilize resources and application services offered by a third-party on the public Internet.
- ***Cloud on Community:***A public cloud is an alliance of multiple infrastructures that are shared among various organizations, allowing for the sharing of applications[5].
- ***Hybrid Clouds:***A hybrid cloud refers to a blend of both private and public cloud computing [5].

The type of cloud service, which provides three models of data modes, is shown in the figure below:

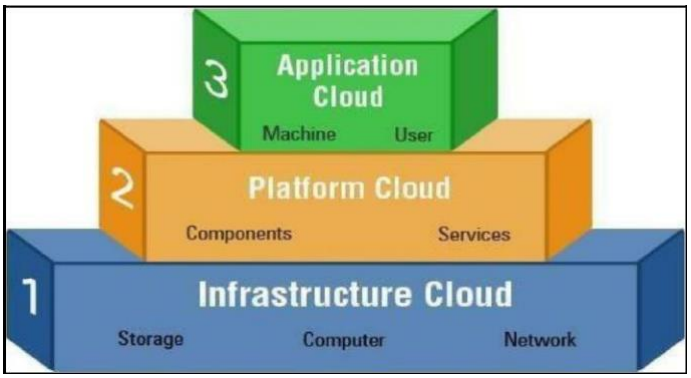

Fig. 1.2: Cloud Service Model

The nature of NIST networks, measured on demand for cloud computing services Self-service, resource pooling and rapid elasticity.

Advantages of cloud computing are as follows:
[6]

- Device & Location Independent
- Secure Storage Management
- It provides a flexible and scalable reliable service.
- 24/7 Support.
- Pay As you Use.
- Reliability, Scalability & Sustainability

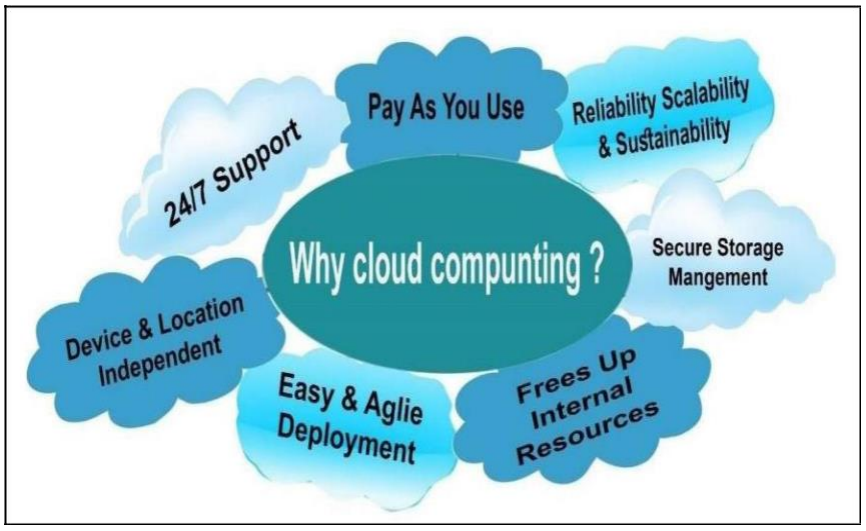

Fig 1.3: Cloud Computing Advantages

**1.1 Cloud Components**

Cloud components refer to various platforms, including the front-end, back-end, delivery dependent on cloud services, and the underlying network.

Therefore, cloud computing components can be broadly categorized as client systems, data center and distributed servers. An exact point of any element plays a definite role [7].

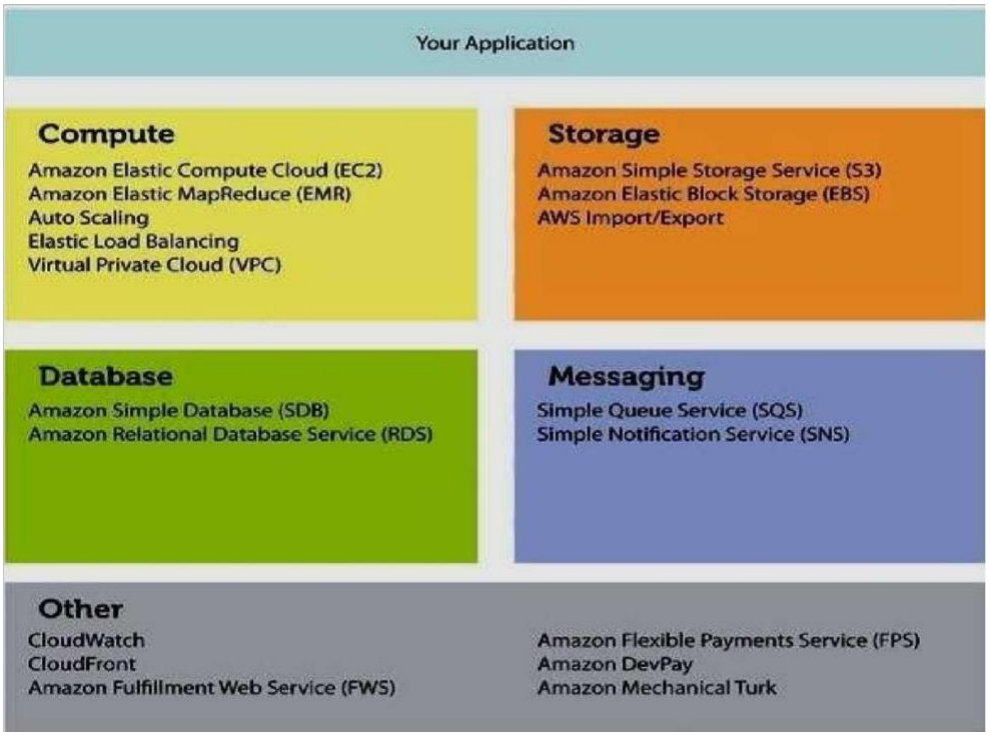


Fig. 1.4: Cloud Computing Components

### *A. Clients*

Cloud Computing has clients for operating Local Area Networks tasks. They are the machines in form of desktop, laptops, tablets etc. Clients‟ main responsibility is to interact in order to manage data on cloud servers [8].

### *B. Data Center*

A collection of various applications of servers build data center. The dataset end user is connected to subscribe to various applications. Datacenter customers may be present over large distances. Now, in a system deployment program, use a technique called virtualization which allows multiple instances of virtual server applications. [9]

### *C. Distributed Servers*

The host is part of a distributed server in the Internet cloud for various applications. But when using the application for the cloud, you will see that your machine will get access to this feature. [10]

# 2. LITERATURE REVIEW

## 2.1 The Management of Resource Assignment in Cloud Computing

The growing concern regarding energy consumption in data centers has led to a rise in interest in the variable management of virtual machines (VMs) on servers. In utility environments where application requirements are dynamic and variable, effective resource management requires capacity planning and demand prediction as essential tools.

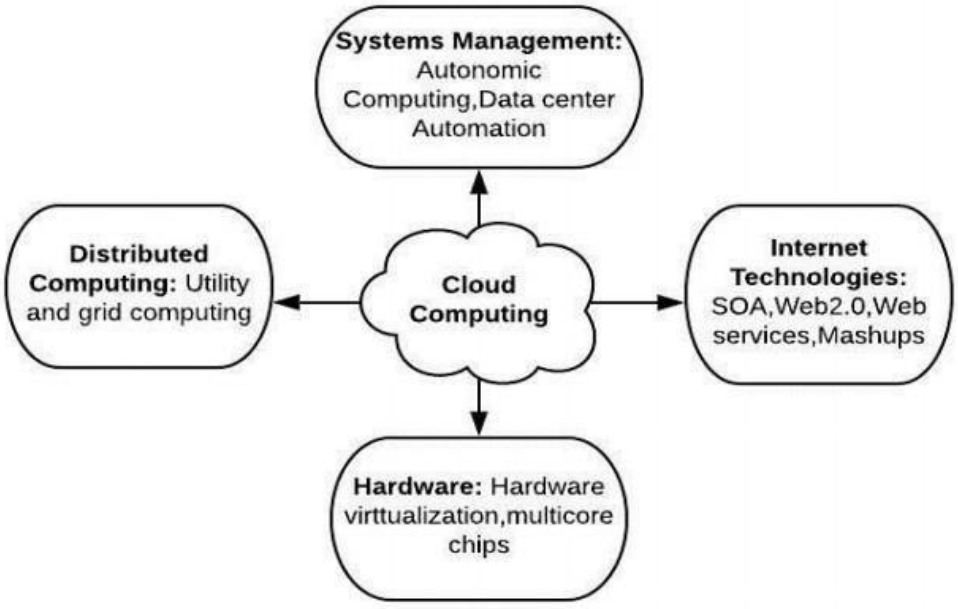
Systems Management: Autonomic Computing,Data center Automation
Distributed Computing: Utility and grid computing
Cloud Computing
Internet Technologies: SOA,Web2.0,Web services,Mashups
Hardware: Hardware virttualization,multicore chips

Fig. 1.5: Advent of Cloud Computing

In Figure 1.5, the primary advancements of cloud computing are displayed. The allocation of resources in cloud computing will be explored in this chapter, with a particular focus on hardware and system management. These elements highlight the potential of cloud resource management to deliver computing services with speed and reliability while also offering benefits in terms of resource utilization and scalability, which can translate to revenue for businesses.

To dynamically consolidate resources from virtual machines (VMs) to physical machines at regular intervals is mandatory in order to effectively manage service level agreements (SLAs) and power consumption in data centers[1]. This approach allows idle machines to be turned off or put into a low-power state to reduce energy usage, while also avoiding overloaded and overheated conditions. The literature presents multiple virtual infrastructure (VI) managers that utilize dynamic resource distribution structures to continuously monitor resource consumption in data centers and reallocate resources according to end user demand and available resources.

## 2.2 Significance of Allocating Resources

Cloud computing is a utility service in which clients consume and pay for services based on usage. This emerging research infrastructure is based on hypervisor, grid computing, and service-oriented architecture (SOA). Cloud computing furnish Infrastructure as a Service (IaaS) in a usage based model that provides personalized resources according to users' requirements. Resource division, in the cloud, is the process of providing at hand resources to the needs hosted by clients. Poorly managed or inadequate resource allocation can cause service disruptions. The resource provisioning manager enables the service provider to allocate resources for each component. For this he makes various resource allocation strategies.

## 2.3 Approaches to Allocating Resources

Resource allocation strategies [2] refer to mechanisms that ensure cloud users receive guaranteed allocation of virtual and/or physical resources with least resource contention. In this process over- or under-provisioning conditions, contentions, scarcity, fragmentation of resources, etc. are kept in mind to avoid. Successful resource allocation requires determining the volume and sort of resources needed by applications to fulfill user tasks, as well as considering the timing of resource allocations and their implications.

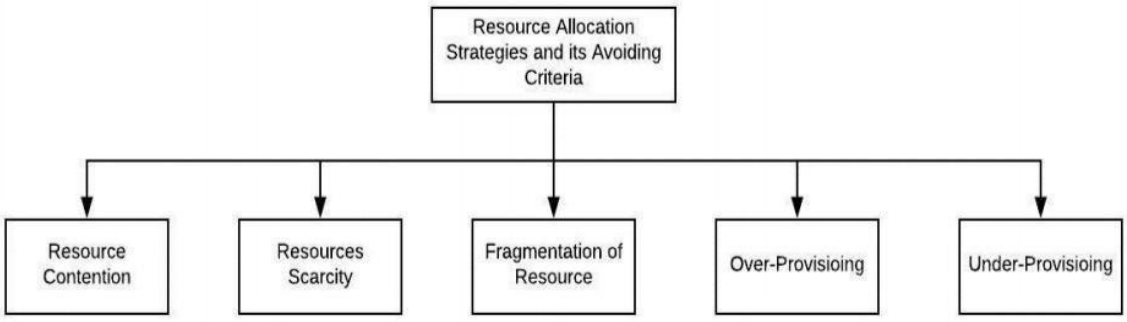


Fig. 1.6: Approaches for Resource Allocation

Let us consider these avoidance criteria individually

- Resource contention occurs when two or more applications attempt to access the same resource simultaneously.

- Inadequate resources result in a shortage of resources.
- Resource fragmentation occurs when resources are available but not arranged in a useful manner. (There may be enough resources, but they may not be intelligently allocated to the demanding application.)
- Provisioning of resources exceeds when more resources are allocated to a task than it actually needs.
- Provisioning of resources lacks when less resources are allocated to a task than it actually needs.

The field of cloud computing has seen a significant amount of research focused on tasks scheduling, resource provisioning, VMs and compute, load balancing , resource management, energy management etc. Among these, load balancing is a focus of researchers just because it is unavoidable aspect of cloud computing for both Cloud Service Providers and Consumers. However, existing literature reviews suggest that one of the reasons for a lack of progress in this area is the absence of a proper classification system for different approaches.

## 2.4 Noteworthy Contribution

This section provides a intellectual overview of the available text on the topic :

Ghomi et al. [11] conducted load balancing algorithms survey in cloud computing. The study categorized these algorithms into seven distinct groups, such as Agent-based load balancing, Natural phenomena-based load balancing and Hadoop-map reduce load balancing,
Milani et al. [12] also studied present load balancing techniques and categorized them into static, dynamic, and hybrid domains. Both studies used proper search operations for retrievingappropriate content from various produced sources and addressed key concerns and challenges in load balancing. However, their analyses were limited to a few QoS metrics, including the time of response,makespan, resource utilization, scalability, shifting time, energy saving and throughput. This survey aims to bridge this gap by considering other necessaryQoS metrics, such as cost of migration, service level violations, balance degree, and ratio of task rejection.

Kalra and Singh [13] performed a comparison between cataloguing algorithms in cloud-based grid computing using five main meta-heuristic techniques, which are Genetic Algorithm (GA), Particle Swarm Optimization (PSO), League Championship Algorithm (LCA), Ant Colony Optimization (ACO), and BAT algorithm. Although the authors conducted a comprehensive evaluation of various methods, yet the study is narrowed to cataloguing algorithms using meta-heuristic methods. Furthermore, their survey solely focuses on evolutionary algorithms, and it lacks a broader categorization.

Mesbahi and Rahmani [14] categorized algorithms of load balancing into three groups

- Based on general algorithm
- Based on architectural
- Based on artificial intelligence-based

They investigated the fundamental needs and necessities for developing and executing a perfect load balancer for a cloud source. Various previous studies uses static as well as dynamic categorizations as basic classification, so as this research. However, the researchers also proposed significant difficulties in designing load balancing algorithms. Moreover, they concluded that the algorithms are most desirable thatpossessqualities of being dynamic, distributive, and non-cooperative.

Kanakala et al. [15] presented a study that distinguished load balancing algorithms as static and dynamic, so as previously debated in other studies. The authors also picked out various challenges in load balancing, such as how to distribute the nodes geographically, transfer time, system performance, hoe to manage and secure the energy. The same has been discussed in the literature for some time. The authors compared existing load balancing algorithms based on several QoS metrics, such as throughput, speed, response time, and migration time. Though there are vast array of algorithms, out of which only eight load balancing algorithms were compared. This is the limitation of study.

Shah et al. [16] presented an overview of load balancing algorithms and classified them basedon the state of the system as static or dynamic, and based on the uniformity of virtual machinetypes as homogeneous or heterogeneous.

In addition, different performance measurements were used to segregate the algorithms and the authors discussed their advantages and disadvantages. However, the paper lacked a systematic outcome of the literature.

Neghabi et al. [17] reviewed load balancing techniques in software-defined networks systematically, arranging their approaches as deterministic and non-deterministic and analyzing its measurements deeply. The authors pose high rise questions and provide answers regarding the role of load balancing, its metric analysis, and challengesof load balancing, importance of load balancing in a network which is software-defined. On the one hand, they present the various advantages in detail, and on the other the limitations of present literature in communication networks and establish a strong foundation for load balancing metrics, although the study does not focus specifically on cloud computing. The paper employs a single-level classification instead of a hierarchical one.

Based on the previously mentioned survey papers, it can be inferred that the current classification systems used in these surveys are lacking. The current methods do not take into account the generalization and specialization characteristics of load balancing algorithms, leading to inadequate conclusions. Additionally, these surveys fail to define difficult intricacies of load balancing algorithms and some metrics of load balancing as frequency of occurrence. Moreover, the available surveys lack a descriptivetext of QoS metric sets, and should be included with few budding points such as migration cost, service level violations, degree of balance, and task rejection ratio. To provide future researchers with insights into developing load balancing algorithms which should be more reliable, improved, perseverant metrics; a taxonomy-based classification system is initiated in this paper. The taxonomy based classification includes various characteristics of load balancing algorithms as nature, state, attribute used for load balancing, mode of execution, type, functionality, and the technique. Performance and economic metrics are also proposed for the classification of QoS metrics.

## 3. PROBLEM STATEMENT

The issue of load unbalancing was examined using the Constructive Generic Framework (CGF) approach, that divide itinto sub-processes. These elements are associated with load balancing - factors, variables, and parameters. In cloud computing, resource scheduling techniques are necessary to manage resources and ensure the cloud service provider generates profits.

The existing algorithms for load balancing can be categorized as static, dynamic, composite, and prioritized. These algorithms aim to minimize response time and boost resource utilization.

## 4. RESEARCH OBJECTIVE

The proposed research/thesis seeks to accomplish the following objectives:

1. Identify the resource allocation and scheduling policies/algorithms.
2. Comparative analysis of mostly used resource allocation and scheduling algorithms.
3. Developing an improved algorithm for resources allocation and scheduling in cloud computing.

## 5. RESEARCH METHODOLOGY

Load balance in the cloud is to properly distribute the workload to virtual machines in a process for proper use. This survey is based on how differently a load balancing algorithm has different metrics. Spending is the lowest possible cost, which helps to allocate resources to balance resource utilization and user satisfaction at load to overcome a problem we find, prompting them to balance and work.

Efficient Load Balancing algorithms for an integrated resource planning and cloud services required for well-organized cloud computing inorder to provide better efficiency. Hence the need for load balancing system that is more efficient for data center cloud & a method for constructing a model of multi resource scheduling to achieve efficiency of resource scheduling in cloud infrastructure. Effective and reasonable load when optimizing algorithms using multidimensional line of dynamic load is achieved by

Selecting the wide use class request of virtual machines through balancing.

Presently Cloud computing is a crucial utility that provides multitenant usage of available resources via Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS). CC aims to provide resources to its end-users without hassle management and perfect allocation mechanism of resources that avoids over-provisioning and under-provisioning to provide elasticity. Therefore, dynamic scheduling algorithms are essential for cloud workload scheduling techniques rather than static scheduling algorithms. Particle Swarm Optimization (PSO) algorithms can be used to calculate the dynamicity of available resource. It also checkthe position of allocated tasks. Resources should be allocated as per Service Level Agreements (SLA) between end-users and Cloud Service Providers (CSP) during service request.While offering resources SLA should be in deal with all its policies and parameters. Proper utilization of resources can boost CSP revenue.

- To recognize the significance and requirement of resource allocation in cloud computing.
- To recognize the diverse policies employed for resource allocation.
- To recognize the different scheduling algorithms created for resource allocation.
- To identify the performance standards of scheduling algorithms utilized for resource allocation.

This study will make use of secondary data also which can be collected from infinite number of resources available for secondary data research that can be used to obtain information which can be collected from periodicals, articles, journals (both manual and electronic), financial reports, and internal organizational documents.

## 6. SCOPE OF RESEARCH

"Cloud computing" includes virtualization, cloud storage, networking, applications, and

online services. A cloud comprises various components, including clients, data centers, and allotted servers, which necessitate high availability, scalability, fault tolerance,efficiency, decreased overhead for customers, less ownership costs, and desired services, among other factors. A perfect load balancing algorithm can only deal with these issues. Such algorithm can only handle various types of loads. This research subjects to create anefficient load balancing algorithm using the Divisible Load Scheduling theorem. The target is to managethroughput and latency, various size clouds etc. performance metrics as per the application requirement. To enhance the efficiency of applications running on virtual machines, it is crucial to allot the correct amount of load over the servers dynamically in the cloud. In this study, we present an improved algorithm that dynamically allocates resources as per precedence and distributes the load across the servers. The experiment was conducted on the Cloud Platform, and the time of reaction was used as a parameter. The results indicate that the proposed algorithm, which uses scaling and transfers,upgrades the executions of applications running in virtual machines.

The management of resources is determining aspect of cloud computing, which also aims to maximize the profits of cloud service providers. To achieve this, scheduling techniques are required, which can be categorized into application layer scheduling, virtualization scheduling, and deployment layer scheduling based on the three cloud services (SaaS, PaaS, and IaaS). The scheduling criteria for each layer should consider the current workload, scheduling objectives, and viewpoints from the cloud user, system, and provider. Genetic algorithms, Ant colony optimization, and particle swarm optimizationlike evolutionary computing algorithmsareplanned for scheduling. However, different scheduling phases like real-time, large scale and dynamic adaption; have challenges. These need to be addressed with the process of complex data and the Internet of Things (IoT) in cloud computing.

The concept of cloud computing involves virtualized computer resources, which has promptedresearch into green cloud computing as a future development trend. The issue of reducingenergy consumption in cloud computing, especially for large-scale clouds, has becomeincreasinglyimportant.

This study discusses various algorithms for the resource allocation.

Here Study and development of efficient algorithms for resource allocation, energy consumption in VM based on NFV and ICO, and VM allocation techniques based on SLA &QoS using CloudSim will carried out.

## 7. RESEARCH SCHEDULE

The following table shows the methodology schedule to achieve research objectives

| SN | Activity | Year 1 | | Year 2 | | Year 3 | |
|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 |
| 1 | □ Primary research assessment<br>□ finding research gap<br>□ Formulation of research problem and objectives | | | | | | |
| 2 | Deep literature review | | | | | | |
| 3 | Research Design & Methodology | | | | | | |
| 4 | Implement the proposed system | | | | | | |
| 5 | □ Testing the result<br>□ Performance analysis | | | | | | |
| 6 | Writing Thesis and preparing the presentation | | | | | | |

## CONCLUSION

Finally, we will evaluate our system and compare the performance with other methods and discuss our future work to improve the proposed system. The expected outcome of this study is a software prototype that will capture the video of the user using a camera

and mark the human face with a rectangle. To develop anextensive comprehension of the load balancing process and its causes, a research methodology was adopted. A literature review was performed to recognize the different methodologies and theories,paradigms, algorithms, and approaches employed to tackle the problem of load unbalancing. The study utilized the Constructive Generic Framework (CGF) methodology [18], which deconstructs the load unbalancing issue into its sub-processes, comprising the corresponding variables, factorsand parameters that affect load balancing.

Here some study part is carried out by reading articles of various researchers about various algorithms (static, dynamic, composite, and prioritized) for the load balancing. The capping goal of these algorithms is to decrease response time and optimize the usage of the resource.

## REFERENCES


[1]. Patel, P., Ranabahu, A. H., &Sheth, A. P. (2009). Service level agreement in cloud computing.

[2]. An, B., Lesser, V., Irwin, D., & Zink, M. (2010, May). Automated negotiation with decommitment for dynamic resource allocation in cloud computing. *In Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems*: Volume 1 (pp. 981-988). International Foundation for Autonomous Agents and Multiagent Systems.

[3]. MArmbrust, A Fox, R Griffith, AD Joseph, R Katz, A Konwinski, G Lee, D Patterson, ARabkin, I Stoica, M Zaharia, A view of cloud computing. Commun. ACM 53(4), 50–58(2009)

[4]. C-J Huang, C-T Guan, H-M Chen, Y-W Wang, S-C Chang, C-Y Li, C-H Weng, Anadaptive resource management scheme in cloud computing. Eng. Appl. Artif. Intell. 26(1),382–389 (2013)

[5]. GS Mukherjee, Mathematical model of cloud computing framework using fuzzy beecolony optimization technique, in *Proceedings of the 2009 International Conference onAdvances in Computing, Control, and Telecommunication*

*Technologies* (IEEE, Piscataway,2009), pp. 664–668

[6]. R Buyya, A Beloglazov, J Abawajy, Energy-efficient management of data center resourcesfor cloud computing: a vision, architectural elements, and open challenges, in *Proceedings ofthe 2010 International Conference on Parallel and Distributed Processing Techniques andApplications* (CSREA, Las Vegas, 2010)

[7]. L Rodero-Merino, L Vaquero, V Gil, F Galan, J Fontan, R Montero, ILlorente, Frominfrastructure delivery to service management in clouds. Futur. Gener. Comput. Syst. 26(8),1226–1240 (2010)

[8]. R Buyya, CS Yeo, S Venugopal, J Broberg, I Brandic, Cloud computing and emerging ITplatforms: vision, hype, and reality for delivering computing as the 5th utility. Futur. Gener.Comput. Syst. 25(6), 599–616 (2009)

[9]. A Beloglazov, R Buyya, Y Lee, AY Zomaya, A taxonomy and survey of energy-efficientdata centers and cloud computing systems. Adv. Comput. 82(1), 47–111 (2011)

[10]. S Khan, A self-adaptive weighted sum technique for the joint optimization ofperformance and power consumption in data centers, in *Proceedings of the 22$^{nd}$International Conference on Parallel and Distributed Computing and CommunicationSystems* (ISCA PDCCS, Kentucky, 2009), pp. 13–18.

[11]. Ghomi EJ, RahmaniAM, Qader NN (2017) Load-balancing algorithms in cloud computing: a survey. J NetwComputAppl 88:50–71

[12]. Milani AS, Navimipour NJ (2016) Load balancing mechanisms and techniques in the cloud environments: systematic literature review and future trends. J NetwComputAppl 71:86–98

[13]. Kalra M, Singh S (2015) A review of metaheuristic scheduling techniques in cloud computing. Egypt Inform J 16(3):275–295

[14]. Mesbahi M, RahmaniAM (2016) Load balancing in cloud computing: a state of the art survey. Int J Mod Educ Comp Sci 8(3):64

[15]. Kanakala VR, Reddy VK, Karthik K (2015, March) Performance analysis of load balancing techniques in cloud computing environment. In: 2015 IEEE

[16]. International Conference on Electrical, Computer and Communication Technologies (ICECCT), pp 1–6

[17]. Shah JM, Kotecha K, Pandya S, Choksi DB, Joshi N (2017, May) Load balancing in cloud computing: methodological survey on different types of algorithm. In: 2017 International Conference on Trends in Electronics and Informatics (ICEI), pp 100–107

[18]. Neghabi AA, Navimipour NJ, Hosseinzadeh M, Rezaee A (2018) Load balancing mechanisms in the software defined networks: a systematic and comprehensive review of the literature. IEEE Access 6:14159–14178

[19]. Afzal S, Kavitha G A taxonomic classification of load balancing metrics: a systematic review

[20]. Vacca JR (2009) Computer and information security handbook. Morgan Kauffman, Burlington, MA, p 208

[21]. Kitchenham B (2004) Procedures for performing systematic reviews. Keele, UK, Keele University, 33(2004), 1–26

[22]. L Chiaraviglio, I Matta, GreenCoop: cooperative green routing with energy-efficientservers, in *Proceedings of the 1st ACM International Conference on Energy-EfficientComputing and Networking, e-Energy* (ACM, New York, 2010),pp. 191–194

[23]. MY Kim, YH Lee, MIP models and hybrid algorithm for minimizing the makespanofparallel machines scheduling problem with a single server. Comput. Oper. Res. 39(1), 2457–2468 (2012)

[24]. D Kliazovich, P Bouvry, SU Khan, Dens: data center energy-efficient network-awarescheduling, in *Proceedings of the ACM/IEEE International Conference on Green Computingand Communications, GreenCom*(IEEE, Piscataway, 2010), pp.69–75.

[25]. Duong-Ba TH, Nguyen T, Bose B, Tran TT. A dynamic virtual machine placement and migration scheme for data centers. IEEE Transactions on Services Computing. 2018:1-11.

[26]. Pires FL, Barán B. A Virtual machine placement taxonomy. In 15th IEEE/ACM international symposium on cluster, cloud and grid computing 2015 (pp. 159-68). IEEE.

[27]. Jonas E, Schleier-Smith J, Sreekanti V, Tsai CC, KhandelwalA,Pu Q, et al.

[28]. Cloud programming simplified: a berkeley view on serverless computing. arXiv preprint arXiv:1902.03383. 2019.

[29]. Choudhary A, Rana S, Matahai KJ. A critical analysis of energy efficient virtual machine placement

[30]. techniques and its optimization in a cloud computing environment. Procedia Computer Science. 2016; 78:132-8.

[31]. ReenaPanwarBhawnaMallick, "A Comparative Study of Load Balancing Algorithms in Cloud Computing", International Journal of Computer Applications (0975 – 8887) Volume 117 – No. 24, May 2015.

[32]. MamtaKhanchi, Sanjay Tyagi, "An Efficient Algorithm for Load Balancing in Cloud Computing" International Journal of Engineering Sciences & Research Technology, June, 2016.